(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENTAMT

(12) Offenlegungsschrift

(11) DE 3416238 A1

(51) Int. Cl. 3: G 10 L 1/00

(21) Aktenzeichen: P 34 16 238.0
(22) Anmeldetag: 2. 5. 84
(43) Offenlegungstag: 20. 12. 84

4,707,858

DE 3416238 A1

(30) Unionspriorität: (32) (33) (31)
02.05.83 US 490701

(71) Anmelder:
Motorola, Inc., Schaumburg, Ill., US

(74) Vertreter:
Grünecker, A., Dipl.-Ing.; Kinkeldey, H., Dipl.-Ing. Dr.-Ing.; Stockmair, W., Dipl.-Ing. Dr.-Ing. Ae.E. Cal Tech; Schumann, K., Dipl.-Phys. Dr.rer.nat.; Jakob, P., Dipl.-Ing.; Bezold, G., Dipl.-Chem. Dr.rer.nat.; Meister, W., Dipl.-Ing.; Hilgers, H., Dipl.-Ing.; Meyer-Plath, H., Dipl.-Ing. Dr.-Ing., Pat.-Anw., 8000 München

(72) Erfinder:
Fette, Bruce A., W. Del Campo Mesa, Ariz., US

(54) Extremschmalband-Übertragungssystem

Ein Übertragungssystem an dessen Enden jeweils eine Vorrichtung zum Analysieren menschlicher Sprache und zum Vergleichen jedes Wortes mit vorgespeicherten Wörtern zur Wort- und Sprechererkennung vorgesehen ist, wobei die Nachricht dann mit charakteristischen Eigenschaften der Stimme des Sprechers digitalisiert wird und ein Signal zur Übertragung mit einer Geschwindigkeit von etwa 75 Bit pro Sekunde gebildet wird, und eine Übertragung der digitalisierten Nachricht zu einem entfernten Terminal erfolgt, das diese Nachricht in eine gesprochene Nachricht in der syntetisierten Stimme des ursprünglichen Sprechers umwandelt.

DE 3416238 A1

GRÜNECKER, KINKELDEY, STOCKMAIR & PARTNER

PATENTANWÄLTE
EUROPEAN PATENT ATTORNEYS

A. GRÜNECKER, DIPL.-ING.
DR. H. KINKELDEY, DIPL.-ING.
DR. W. STOCKMAIR, DIPL.-ING., A.E.S. (CALTECH)
DR. K. SCHUMANN, DIPL.-PHYS.
P. H. JAKOB, DIPL.-ING.
DR. G. BEZOLD, DIPL.-CHEM.
W. MEISTER, DIPL.-ING.
H. HILGERS, DIPL.-ING.
DR. H. MEYER-PLATH, DIPL.-ING.

8000 MÜNCHEN 22
MAXIMILIANSTRASSE 58

30. April 1984
P 18 697

MOTOROLA, INC.
1303 E. Algonquin Road, Schaumburg,
Illinois 60196, USA

Extremschmalband-Übertragungssystem

P a t e n t a n s p r ü c h e

1. Extremschmalband-Übertragungssystem mit einem Wandler zum Umwandeln menschlicher Sprache in elektrische Signale, g e k e n n z e i c h n e t durch:

eine Analysiervorrichtung (15), die elektrische Signale vom Wandler (14) empfängt und eine Vielzahl von Signalen abgibt, die eine Vielzahl von Eigenschaften darstellen, die eine menschliche Stimme charakteri-

sieren;

eine Speichervorrichtung (20) in der Signale gespeichert sind, die eine Vielzahl gesprochener Wörter darstellen;

eine Worterkennungsvorrichtung (16), die mit der Analysiervorrichtung (15) und mit der Speichervorrichtung (20) zum Empfang von zumindest eines Teiles der Vielzahl von Signalen zum Vergleichen des empfangenen Teiles der Vielzahl von Signalen mit den gespeicherten Signalen verbunden ist, und Signale abgibt, die speziell gesprochene Wörter darstellen, und

eine Digitalumwandlungsvorrichtung, die mit der Worterkennungsvorrichtung (16) zum Empfang der spezifisch gesprochenen Wörter darstellenden Signale verbunden ist, zum Umwandeln der empfangenen Signale in eine Digitalform mit einer Geschwindigkeit von weniger als 300 Bit pro Sekunde.

2. Extremschmalband-Übertragungssystem nach Ansrpuch 1, dadurch gekennzeichnet, daß die Analysiervorrichtung (15) eine Analysierschaltung (32) für eine linear vorhersagbare Codierung aufweist.

3. Extremschmalband-Übertragungssystem nach Anspruch 1 oder 2, dadurch gekennzeichnet, daß die Worterkennungsvorrichtung (16) eine Vorrichtung (42, 43, 45) zum Erkennen des Beginns und des Endes eines gesprochenen Wortes aufweist.

4. Extremschmalband-Übertragungssystem nach einem der Ansprüche 1 bis 3, dadurch gekennzeichnet, daß die Speichervorrichtung (20) Signale gespeichert hat, die eine Vielzahl von Wörtern darstellen, die von einer Vielzahl unterschiedlicher Induviduen gesprochen wurden, und daß eine Sprechererkennungsvorrichtung (18)

mit der Speichervorrichtung (20) und der Analysiervorrichtung (15) vorgesehen ist, die zumindest einen Teil der Vielzahl von Signalen von der Analysiervorrichtung (15) empfängt, die empfangenen Signale mit den gespeicherten Signalen vergleicht und Signale abgibt, die spezielle Wörter darstellen, die von einer spezifischen, der unterschiedlichen Individuen gesprochen wurden.

5. Extremschmalband-Übertragungssystem nach Anspruch 4, dadurch gekennzeichnet, daß die Sprechererkennungsvorrichtung (18) eine Schaltung zum Modifizieren gespeicherter Wörter eines Individium nach einer Sprechererkennung aufweist.

6. Extremschmalband-Übertragungssystem nach Anspruch 5, dadurch gekennzeichnet, daß die Analysiervorrichtung (15) eine Schaltung (32) zum Erhalten von LPC-Koeffizienten einer linearen vorhersagbaren Codierung und die Sprechererkennungsvorrichtung (18) eine Schaltung zur Mittelwertbildung der LPC-Koeffizienten aufweist.

7. Extremschmalband-Übertragungssystem nach Anspruch 6, dadurch gekennzeichnet, daß die Sprechererkennungsvorrichtung (18) eine Schaltung (20) zum Zurückstellen einer Entscheidung bezüglich der Identität des Sprechers aufweist, wenn er der Vergleich eines gesprochenen Wortes mit gespeicherten Signalen, die eine Vielzahl von einer Vielzahl von unterschiedlichen Induviduen gesprochenen Wörtern darstellen, innerhalb eines vorbestimmten Unsicherheitsbereichs liegt.

8. Extremschmalband-Übertragungssystem nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß die Signalumwandungsvorrichtung (20) eine Einrichtung zum Umwandeln von Buchstaben jedes spezifischen gesprochenen Wortes in ASCII Digi-

talcodierungen für eine Übertragung.

9. Extremschmalband-Übertragungssystem nach einem der Ansprüche 4 bis 7, gekennzeichnet durch eine Nachrichtenformatierungsvorrichtung (20), die mit der Sprechererkennungsvorrichtung (18) und der Worterkennungsvorrichtung (16) zum Formatieren jeder in den Wandler gesprochenen Nachricht in ein elektrisches Digitalsignal verbunden ist, das eine Vielzahl von Bits enthält, die die Nachricht darstellen, sowie eine Vielzahl von Bits, die den Sprecher charakterisieren.

10. Extremschmalband-Übertragungssystem nach Anspruch 9, dadurch gekennzeichnet, daß die Nachrichtformatierungsvorrichtung auch Bits verarbeitet, die Eigenschaften darstellen, die die Stimme des Sprechers charakterisieren.

11. Extremschmalband-Übertragungssystem nach Anspruch 10, gekennzeichnet, durch eine Vorrichtung (20) zum Übertragen der Digitalsignale von der Umwandlungsvorrichtung zu einer entfernt angeordneten Einheit (12), eine Vorrichtung (20) zum Empfangen von Digitalsignalen und einer Synthetisiervorrichtung (22) zum Umwandeln der Digitalsignale in die synthetisierte menschliche Sprache, die charakteristisch für die Stimme des Sprechers ist.

12. Extremschmalband-Übertragungsverfahren, gekennzeichnet durch die Schritte

Umwandeln von menschlicher Sprache in elektrische Signale,

Analysieren der elektrischen Signale, um eine Vielzahl von Signalen abzugeben, die eine Vielzahl von Eigenschaften darstellen, die eine menschliche Stimme charakterisieren,

Speichern von Signalen, die eine Vielzahl von gesprochenen Wörtern darstellen,

Vergleichen zumindest einiger der Vielzahl von Signalen mit den gespeicherten Signalen, um spezifische Wörter in der menschlichen Sprache zu bestimmen und Signale abzugeben, die die spezifischen Wörter darstellen, und

Umwandeln der abgegebenen Signale, die spezifische Wörter darstellen in eine Digitalform mit einer Geschwindigkeit geringer als 300 Bits pro Sekunde.

13. Verfahren nach Anspruch 12, g e k e n n z e i c h n e t durch Erkennen des Beginns und des Endes jedes gesprochenen Worts vor dem Vergleichen.

14. Verfahren nach Anspruch 12 oder 13, dadurch g e k e n n z e i c h n e t, daß das Speichern eine Speichern von Signalen umfaßt, die eine Vielzahl von von einer Vielzahl unterschiedlicher Induviduen gesprochenen Wörtern darstellen und daß das Vergleichen das Zuführen von Signalen umfaßt, die repräsentativ sind für das individuelle Sprechen der spezifischen Wörter.

15. Verfahren nach einem der Ansprüche 12 bis 14, dadurch g e k e n n z e i c h n e t, daß beim Analysieren Koeffizienten mit linearer vorhersagbarer Codierung erzeugt und die Koeffizienten vor dem Vergleichen gemittelt werden.

16. Verfahren nach einem der Ansprüche 12 bis 15, dadurch g e k e n n z e i c h n e t, daß das Vergleichen das Zurückstellen einer Entscheidung bezüglich des induviduellen Sprechens umfaßt, wenn der Vergleich eines gesprochenen Wortes mit gespeicherten Signalen, die für eine Vielzahl von durch eine Vielzahl von unterschied-

lichen Induviduen gesprochenen Wörtern sind, innerhalb eines vorbestimmten Unsicherheitsbereichs liegt.

17. Verfahren nach einem der Ansprüche 12 bis 16, gekennzeichnet durch Modifizieren gespeicherter Signale, die eine Vielzahl von von einem Induviduum gesprochenen Wörtern darstellen, nachdem dieses spezielle Induviduumerkannt worden ist, und gemäß der zuletzt gesprochenen Sprache des Individuums.

18. Verfahren nach einem der Ansprüche 12 bis 17, dadurch gekennzeichnet, daß eine Vielzahl vorbestimmter Nachrichten gespeichert und dem Sprecher eine Liste möglicher nächster Wörter nach Erkennung des Endes eines Wortes angezeigt wird.

19. Verfahren nach einem der Ansprüche 12 bis 18, gekennzeichnet durch Formatieren der menschlichen Sprache nach Umwandlung in die Digitalform in ein elektrisches Digitalsignal mit einer Vielzahl von Bits, die eine Nachricht darstellen und eine Vielzahl von Bits, die charakteristische Eigenschaften der menschlichen Stimme darstellen, und Übertragen des elektrischen Digitalsignals zu einem entfernten Terminal.

20. Verfahren nach Anspruch 19, gekennzeichnet durch Empfangen eines von einem entfernten Terminal ausgesandten elektrischen Digitalsignals und Umwandeln des empfangenen Signals in eine gesprochene Nachricht in einem synthetisierten Stimme, die angenähert die charakteristischen Eigenschaften des ursprünglichen Sprechers am entfernten Terminal hat.

3416238

GRÜNECKER, KINKELDEY, STOCKMAIR & PARTNER

PATENTANWÄLTE
EUROPEAN PATENT ATTORNEYS

A. GRÜNECKER, DIPL.-ING.
DR. H. KINKELDEY, DIPL.-ING.
DR. W. STOCKMAIR, DIPL.-ING., A.E.S. (CALTECH)
DR. K. SCHUMANN, DIPL.-PHYS.
P. H. JAKOB, DIPL.-ING.
DR. G. BEZOLD, DIPL.-CHEM.
W. MEISTER, DIPL.-ING.
H. HILGERS, DIPL.-ING.
DR. H. MEYER-PLATH, DIPL.-ING.

8000 MÜNCHEN 22
MAXIMILIANSTRASSE 58

30. April 1984
P 18 697-57/to

MOTOROLA, INC.
1303 E. Algonquin Road, Schaumburg,
Illinois 60196, USA

Extremschmalband-Übertragungssystem

Beschreibung

In Übertragungssystemen ist es äußerst wünschenswert, Nachrichten mittels Sprache auszutauschen. Andererseits ist es erwünscht, digitale Schaltungen zu verwenden, da ein Großteil dieser Schaltungen auf einem einzigen integrierten Schaltungschip untergebracht werden können, was den erforderlichen Raum- und Energiebedarf wesentlich verringert. Digitale Darstellungen der menschlichen Sprache erfordern jedoch im allgemeinen eine verhältnismäßig große Bandbreite, so daß sie für viele Arten von Übertragungsmedien, etwa Telefonleitungen oder dergleichen, nicht geeignet sind. Die Bit-Übertragungsgeschwindigkeit (Bandbreite) von Nachrichten soll deshalb so niedrig wie möglich sein. Unter "Schmalband" wird üblicherweise eine Bit-Übertragungsgeschwindigkeit von etwa 2 000 Bits pro Sekunde verstanden. Bekannte Vorrichtungen arbeiten über

300 Bits pro Sekunde und alles, was darunter liegt, soll als "Extremschmalband" bezeichnet werden.

Die vorliegende Erfindung betrifft ein Extremschmalband-Übertragungssystem und ein Verfahren zum Nachrichtenaustausch in einem extremen Schmalband, wobei menschliche Sprache in elektrische Signale umgewandelt und analysiert wird, so daß sich Signale ergeben, die Eigenschaften darstellen, welche das spezielle menschliche Sprechen charakterisieren. Die Wörter der Nachricht werden dann mit Wörtern in einem Speicher verglichen, so daß das spezielle Wort erkannt wird und falls erwünscht auch der spezielle Sprecher, der dieses Wort ausgesprochen hat. Ein das spezielle Wort darstellendes Digitalsignal, das eine ASCII- oder numerische Kodierung sein kann und die Position des Wortes im Speicher angibt, wird mit Digitalsignalen kombiniert, die die Stimme des Sprechers charakterisieren, damit sich eine Nachricht ergibt mit einer Bit-Geschwindigkeit wesentlich unter 300 Bit pro Sekunde, wobei die Nachricht zu einem entfernten Endgerät übertragen wird. Dieses Endgerät synthetisiert die menschliche Stimme, so daß die Nachricht derart ertönt, als wenn die ursprüngliche Stimme sprechen würde. Verschiedene Verfahren und Einrichtungen dienen dazu, die korrekte Erkennung jedes Wortes und des speziellen Sprechers zu gewährleisten einschließlich einer Mittelwertbildung von LPC-Koeffizienten, Hinausschieben einer Entscheidung bezüglich der Identität des Sprechers, wenn der Vergleich der gesprochenen mit den gespeicherten Wörtern innerhalb eines vorbestimmten Unsicherheitsbereichs liegt und Modifizieren beziehungsweise auf den neuesten Stand Bringen der gespeicherten Wörter eines individuellen Sprechers, nachdem dieser erkannt wurde.

Der Erfindung liegt die Aufgabe zugrunde, ein neues und verbessertes Extremschmalband-Übertragungssystem anzugeben.

Ferner soll ein verbessertes Verfahren des Nachrichtenaustausches mittels Extremschmalband aufgezeigt werden.

An der empfangenden Endstation soll eine Stimme synthetisiert werden, die gleich derjenigen des ursprünglichen Sprechers ist.

Die Erkennung des Sprechers soll äußerst genau erfolgen.

Ein Ausführungsbeispiel der Erfindung wird nachstehend unter Bezugnahme auf die Zeichnung beschrieben. Es zeigen

- Figur 1 ein vereinfachtes Blockschaltbild eines Extremschmalbandnachrichten- oder Übertragungssystems der Ausführungsform der Erfindung,
- Figur 2 ein Blockschaltbild der LPC-Analysiereinheit des Systems nach Figur 1,
- Figur 3 ein Blockschaltbild der CPU-Einheit des Systems nach Figur 1,
- Figur 4 ein Blockschaltbild der Worterkennungseinheit des Systems nach Figur 1,
- Figur 5 ein Blockschaltbild der Synthetisiereinheit des Systems nach Figur 1,
- Figur 6 ein Flußdiagramm zur Veranschaulichung des Beginns und der Beendigung einer Wortidentifikation in der Worterkennungseinheit der Figur 4,
- Figur 7 ein Flußdiagramm beziehungsweise ein Syntaxbaum bestimmt für militärische Zwecke und
- Figur 8 vier typische Anzeigebilder im Zusammenhang mit dem Flußdiagramm der Figur 7.

Figur 1 zeigt das Extremschmalband-Übertragungssystem gemäß dem Ausführungsbeispiel der Erfindung. Ein Ortsterminal 10 und ein entferntes Terminal 12 sind über ein geeignetes Mittel, etwa Telefonleitungen oder dergleichen, ver-

bunden. Das Ortsterminal 10 weist ein Mikrofon 14 zum Umwandeln der menschlichen Sprache in elektrische Signale in üblicher Art auf und ist mit einer LPC-Analysiereinheit 15 und einer Worterkennungseinheit 16 verbunden. LPC-Analyse bedeutet Analyse einer linearen vorhersagbaren Kodierung. Die LPC-Analysiereinheit oder -schaltungsplatte 15 ist an eine zentrale Verarbeitungseinheit CPU 18 angeschlossen, die wiederum mit einem Rechner 20 in Verbindung steht, der ein Tastenfeld, einen Austauschplattenspeicher (Floppydiscspeicher) und eine Sichtanzeige aufweist. Die Worterkennungseinheit 16 ist mit dem Personalrechner 20 und eine Synthetisiereinheit oder -schaltungsplatte 22 ist ebenfalls mit dem Rechner 20 verbunden. Der Ausgang der Synthetisiereinheit 22 liegt an Kopfhörern 23 oder einem anderen Wandler geeigneter Art zum Umwandeln elektrischer Signale von der Synthetisiereinheit 22 in Schall.

Figur 2 zeigt in größerer Einzelheit ein Blockschaltbild der LPC-Analysiereinheit 15 in Form eines vollständigen digitalen Sprachverarbeitungssystems, wie es im einzelnen in der noch schwebenden US-Patentanmeldung mit der Bezeichnung "Digital Voice Processing System" und dem Aktenzeichen 309 640 vom 8. Oktober 1981 beschrieben ist. Die LPC-Analysiereinheit ist nur ein Teil des in Figur 2 veranschaulichten Systems und ist im einzelnen in der US-PS 4 378 469 erläutert. Das vollständige Verarbeitungssystem ist deshalb beschrieben, weil es einen Teil der LPC-Analysiereinheit 15 darstellt und der Synthetisierteil der Einheit 15 zur Synthetisierung der menschlichen Stimme verwendet werden kann, so daß sie am entfernten Terminal 12 wie das Sprechen eines Sprechers ertönt. Im vorliegenden System wird der Synthetisierer der Einheit 15 nicht verwendet. Der Fachmann erkennt jedoch, daß diese Einheit ohne weiteres an Stelle der Synthetisiereinheit 22 eingesetzt werden kann.

Gemäß Figur 2 werden Tonfrequenzsignale von dem Mikrofon 14 über eine AVR-Schaltung 25 mit automatischer Verstärkungsregelung und ein Tiefpaßfilter 26 einer Abtast- und Halteschaltung 28 zugeführt. Diese arbeitet mit einem Analog-/Digitalwandler 30 zusammen, um für jede durch die Abtast- und Halteschaltung 28 durchgeführte Abtastung eine 12 Bit-Digitaldarstellung abzugeben. Diese Digitalwerte von dem A/D-Wandler 30 werden einer LPC-Analysiereinheit 32 zugeführt, die in der vorgenannten Patentschrift im einzelnen beschrieben ist. Die LPC-Analysiereinheit 32 gibt mehrere Signale ab, die unterschiedliche Eigenschaften darstellen, die eine menschliche Stimme charakterisieren, wie den Tonhöhenfrequenzbereich und eine Abschätzung der vokalen Spurlänge sowie wahlweise einsetzbare zusätzliche Eigenschaften, wie die glottale Erregungsform im Frequenzbereich und der Heiserkeitsgrad etc. Die Signale von der LPC-Analysiereinheit 32 umfassen auch einen RMS-Durchschnittswert und eine vorbestimmte Anzahl von LPC-Koeffizienten, nämlich in diesem Ausführungsbeispiel zehn. Alle diese Signale von der LPC-Analysiereinheit 32 werden über eine Schnittstelle 34 der CPU 18 zur Speicherung und Verarbeitung zugeführt. Ein detaillierteres Blockschaltbild der CPU 18 ist in Figur 3 gezeigt. Bei diesem Ausführungsbeispiel ist die CPU 18 die im Handel erhältliche CMT-68K-CPU. Da die in Figur 3 veranschaulichte CPU 18 im Handel erhältlich ist, kennt der Fachmann die Arbeitsweise. Da alle Blöcke ausreichend definiert sind, soll deren Funktion nicht im einzelnen beschrieben werden.

Obwohl die verschiedensten Einrichtungen als Worterkennungseinheit 16 verwendet werden können, kommt bei der vorliegenden Ausführungsform die im Handel erhältliche Einheit VRM102 zum Einsatz, die anhand der Figur 4 erläutert wird. Die Tonfrequenzsignale vom Mikrofon 14 werden an den Audioeingang angelegt und über einen Vorverstärker 35 zum 16 Filter-Analysierer 37 geleitet. Der 16 Filter-Analysierer 37 führt grundsätzlich die Analysierfunk-

tion der LPC-Analysiereinheit durch und der Fachmann erkennt, daß eine Worterkennungseinheit auch auf Signale der LPC-Analysiereinheit 15 basieren kann. Das Ausgangssignal des 16 Filter-Analysierers 37 wird über einen Gleichrichter 39 an einen 8 Bit-Analog-/Digitalwandler 40 angelegt. Dieser A/D-Wandler 40 ist mit einem 6802 Mikroprozessor 42, einem 4K-RAM-Speicher 43 und einem 4K-ROM-Speicher 45 verbunden. Die Worterkennungseinheit 16 besitzt auch mehrere Anschlüsse und Puffer zum Nachrichtenaustausch mit dem Personalrechner 20, dessen Funktion bekannt ist und hier nicht im einzelnen beschrieben wird.

Spektralamplituden des Gleichrichters 39 werden alle 5 ms durch den A/D-Wandler 40 ausgelesen. Das System mißt die Spektraldifferenz zwischen dem augenblicklichen Spektrum und dem Hintergrundrauschen. Überschreitet diese Differenz einen ersten Schwellenwert, dann markiert das System den möglichen Beginn eines Wortes und spektrale Abtastungen werden in dem "UNBEKANNTEN"-Schablonenspeicher 4K-RAM-Speicher 43 aufgezeichnet. Nun wird die Empfindlichkeit auf Spektraländerungen erhöht und neue Spektren werden immer dann aufgezeichnet, wenn eine gegen einen zweiten Schwellenwert gemessene geringfügige Änderung zwischen dem augenblicklichen und dem letzten Spektrum auftritt. Bei jeder signifikanten Änderung wird ein im Personalrechner 20 angeordneter Abtastzähler (NSAMP) aufgezählt. Diese Zählung muß ein Minimum von MINSAM, nämlich 16 unterschiedliche Spektralformen erreichen, bevor das System ein Wort als gültig erklärt, sonst wird der Schall als Hintergrundrauschen bestimmt. Jeder 5 ms-Rahmen, der keine signifikante Spektraländerung aufweist, ist ein Hinweis auf das Wortende. Vergehen 160 ms ohne Spektrumsänderung, dann wird das letzte Spektrum als wahrscheinliches Wortende erklärt und eine Musterübereinstimmungsprüfung beginnt. Ein Flußdiagramm dieses Verfahrens ist in Fig. 6 veranschaulicht.

Der Ablauf beginnt mit einem Zustand 47, der mit "Ruhezustand, kein Wort" bezeichnet ist. Der Abtastzähler (NSAMP) beginnt bei Null zu zählen und wenn die Differenz zwischen dem augenblicklichen Spektrum und dem Hintergrundrauschen den Schwellenwert t1 überschreitet, dann läuft das Verfahren zum Zustand 48, der mit "möglicher Wortbeginn" bezeichnet ist. Überschreitet die Differenz zwischen dem augenblicklichen und dem letzten Spektrum nicht den zweiten Schwellenwert t2, dann geht der Ablauf zum Kreis 49, der mit "NSCNG = NSCHG + 1" bezeichnet ist. Ist die Zeit seit der letzten Spektraländerung kurz, dann kehrt der Ablauf zurück zum Zustand 48, um die Messung von Spektraländerungen zwischen dem augenblicklichen und dem letzten Spektrum fortzusetzen. Ist die Zeit seit der letzten Spektraländerung lang - bei dem vorliegenden Ausführungsbeispiel etwa 160 ms - dann folgt im Ablauf der Zustand 50, der mit "mögliches Wortende" bezeichnet ist. Ist die Zählung in dem Abtastzähler geringer als 16, dann kehrt der Ablauf zurück zum Zustand 47 und beginnt erneut und die Spektraländerungen werden als zu kurz für ein Wort betrachtet, so daß sie Hintergrundrauschen darstellen müssen. Überschreitet die Zählung des Abtastzählers den Wert 16, dann folgt der Zustand 52, mit "Wortende, stelle Übereinstimmung des Musters mit Ausgangswert her". Somit stellt das System fest, daß ein Wort gesprochen wurde und es beginnt die Musterübereinstimmungsprüfung.

Sobald die Spektraländerung zwischen dem augenblicklichen und letzten Spektrum den Schwellenwert t2 überschreitet, folgt Zustand 51, der mit "Bringe signifikantes Spektralmodell auf neuesten Stand" beschrieben ist. Ist der Eingangspuffer des Abtastzählers NSAMP nicht gefüllt, dann kehrt der Ablauf zum Zustand 48 für die nächste 5 ms-Abtastung zurück. Wird der Eingangspuffer des Abtastzählers NSAMP bei einer großen Spektraländerung gefüllt, dann geht der Ablauf direkt zum Zustand 50, wo dies als Wortende bestimmt wird und es folgt Zustand 52, in dem die Herstellung der Musterübereinstimmung beginnt. Wird der Ein-

gangspuffer des Abtastzählers NSAMP aufgrund eines kurzen Wortes nicht gefüllt, dann ergeben sich schließlich keine Spektraländerungen in den Abtastungen und der Ablauf geht zum Zustand 49 wie zuvor beschrieben.

Bei dem Terminal des vorliegenden Ausführungsbeispiels ist eine vorbestimmte Anzahl von Sprechern autorisiert, das Terminal zu verwenden und Beispiele vorbestimmter Wörter und Phrasen, wie sie von jedem Sprecher gesprochen wurden, sind in dem Austauschscheibenspeicher des Rechners 20 gespeichert. Die Worterkennungseinheit 16 dient zur Unterstützung bei der Sprechererkennung bei einer etwas vereinfachten Ausführungsform. Wenn ein spezieller Sprecher auf das System zugreift, identifiziert er sich sprachlich durch Name, Stand und Personalnummer oder irgendeine andere Identifizierungszahl. Der Beginn und das Ende jedes Wortes wird von der Worterkennungseinheit 16 festgestellt, die den Personalrechner 20 von dem gesprochenen Wort in Kenntnis setzt. Eine elektrische Darstellung von LPC-Parameterdaten der LPC-Analysiereinheit 15 wird über den gesprochenen Bereich jedes Wortes gemittelt, dann in der CPU 18 mit einem gespeicherten Beispiel vom Rechner 20 zur Übereinstimmung gebracht. Die Ergebnisse der Übereinstimmungsprüfung werden mit einem Schwellenwert verglichen, um eine Entscheidung über die Identität des Sprechers herbeizuführen.

Während der Benutzer das System weiter verwendet, erkennt der Rechner 20 Stellen in Sätzen, wo die Anzahl möglicher nächster Wörter verhältnismäßig gering ist, wie dies jetzt beschrieben wird. An diesen syntaktischen Knoten lädt der Personalrechner 20 Muster oder Schablonen, d.h. gespeicherte Modelle von Wörtern aller Sprecher für diese nächsten möglichen Wörter. Beim nächsten gesprochenen Wort erkennt die Worterkennungseinheit diese Tatsache und vergleicht die in das System geladenen Muster mit der Darstellung des gerade gesprochenen Wortes. Die Worterkennungs-

einheit zeigt das gesprochene Wort an der Sichtanzeige des Rechners 20 und auch den Sprecher an. Der Rechner 20 besitzt einen Abstimmzähler für jeden der möglichen autorisierten Sprecher. Der Zähler des angezeigten Sprechers wird mit jedem erkannten Wort aufgezählt bis zu einem Maximum von 25 und die Zähler aller nichtangezeigten Sprecher werden abwärts gezählt bis zu einer unteren Grenze von Null. Wird beispielsweise eine Geheiminformation angefordert, dann werden die Zähler geprüft und als identifizierter Sprecher derjenige bestimmt, dessen Zählung über 15 liegt, während alle anderen Zählungen unter 8 liegen müssen. Werden diese Bedingungen nicht erfüllt, dann wird die Geheiminformation abgelehnt. Das System kann den Benutzer im weiteren Identifikationsalgorithmus auffordern, beliebige Wörter zu sprechen, bis ein eindeutiger Gewinner mit entsprechendem Abstand angezeigt wird, oder das System kann in seinem normalen Ablauf fortfahren und zu einem späteren Zeitpunkt die Information nochmals anfordern. Das System kann eine Änderung des Sprechers innerhalb von maximal 10 Wörtern erkennen. Auch ist der Sprecheridentifikationsalgorithmus dem Benutzer im allgemeinen erkennbar und er weiß nicht, daß seine Stimme während des normalen Ablaufs analysiert wird.

Die Verifikationssubsystemsoftware wird von den Austauschplatten des Rechners 20 geladen und dieses Laden wird durch Prüfsummentests verifiziert. Als nächstes werden statistische Muster jedes bekannten Sprechers ebenfalls geladen. Während der unbekannte Sprecher spricht, werden Langzeitstatistiken der LPC-Reflexionskoeffizienten in Echtzeit über die letzten 30 Sekunden der Sprache berechnet. Diese Statistiken schließen eine Mittelwert- und Standardabweichung der Tonhöhe und die ersten 10 Reflexionskoeffizienten ein. Am Ende jedes Wortes, wie es durch die Worterkennungseinheit 16 bestimmt wurde, berechnet die CPU 18 die Mehalanobisabstandsmetrik zwischen dem unbekannten Wort und dem Muster jedes Sprechers. Der Mehalanobis-

abstand gewichtet den Abstand mit der Fähigkeit jedes Messungs-Eigenfektors, um den bekannten Sprecher von der allgemeinen Bevölkerung zu unterscheiden. Schließlich informiert die CPU über den Sprecher mit der besten Übereinstimmung und bestimmt die Genauigkeit der Schätzung durch den Mehalanobisabstand unter Verhältnisbildung zur Standardabweichung dieses Sprechers und durch das Verhältnis zu der nächstbesten Übereinstimmung. Zweideutige Ergebnisse d.h., wenn die Übereinstimmung innerhalb eines vorbestimmten Unsicherheitsbereichs liegt, bewirken, daß das System eine Entscheidung zurückstellt, wodurch die Genauigkeit erhöht wird. Schließlich wird am Ende des Nachrichtenaustausches dem Sprecher die Möglichkeit gegeben, sein Stimmenmodell durch die zusammengesetzten Statistiken dieses Nachrichtenaustausches auf den neuesten Stand zu bringen.

Die LPC-Analysiereinheit 15 und die CPU 18 besitzen auch eine Trainingsarbeitsweise bei der sich diese Statistiken eines gegebenen Sprechers ergeben und in der die Eigenfektoren und Werte des Modells dieses Sprechers berechnet werden. Das System kann diese Daten zur Speicherung auf den Austauschplatten des Rechners 20 aufwärts laden. Während die Worterkennungseinheit 16 als getrennte Einheit des Systems veranschaulicht wird, weis der Fachmann, daß sie in einfacher Weise auch in die LPC-Analysiereinheit 15 oder die CPU 18 eingefügt sein kann, so daß diese Einheiten die Aufgaben der Erkennung des Beginns und Endes eines Wortes, des spezifischen Wortes und des Sprechers durchführen können. Auch können Schablonen oder Wortmodelle, die allgemein repräsentativ für jedes spezielle zu erkennende Wort sind, an Stelle eines Wortmodells für jedes von jedem Sprecher gesprochene zu erkennende Wort verwendet werden, wobei nur die speziellen Wörter durch die Einrichtung erkannt würden und nicht jedoch jeder spezielle Sprecher.

Ein typisches Beispiel einer militärischen Verwendung des

vorliegenden Systems sei nun in Verbindung mit den Fig. 7 und 8 erläutert. Bei dieser speziellen Ausführungsform ist das System so aufgebaut, daß es den Verwender mit einbezieht, ein geographisches Truppenmodell, Nachschub und geographische Umwelt auf den neuesten Stand zu bringen. Bei der grundsätzlichen Situation dieses Ausführungsbeispiels fordert der Benutzer Information von dem Terminal an und, falls er richtig erkannt und geprüft wurde, wird die Information von einer entfernten Quelle gegeben. Es sei für dieses spezielle Ausführungsbeispiel angenommen, daß das System um einen halben Bildschirm nach links, rechts oben oder unten schwenken kann, oder nach Norden, Süden Osten oder Westen bei n-Meilen. Es soll ferner die Fähigkeit besitzen, eine Fokusierte oder eine breitere Darstellung zu bieten, und zeigt wesentliche geographische Merkmale, etwa eines eines Landesstaates einer Stadt von Grenzen, Straßen oder Hügel an. Bei der speziellen Anwendung des Systems werden 55 Wörter und ein Syntaxnetzwerk mit semantischen Zuordnungen zu jedem Knoten des Netzwerks verwendet, wie dies Fig. 7 veranschaulicht. Ein Syntaxnetzwerk leitet interaktiv die Auswahl von möglichen, nächsten Wörtern von allen dem System bekannten Wörtern im Kontext aller Sätze, die das System versteht. Der Sprecher kann jederzeit sagen "Löschen" um einen neuen Satz zu beginnen, oder er kann sagen "Auslöschen" um in einem Satz ein Wort zu ersetzen. Wörter wie "UH, THE",Atemgeräusch und Zungenschlagen sind Modellwörter, die gespeichert werden und die von dem System absichtlich ignoriert werden. Das System hilft dem Benutzer interaktiv, wenn dieser spricht. Erwartet das System von ihm, daß er einen Satz beginnt, d.h., wenn die Worterkennungseinheit 16 den Anfang eines ersten Wortes feststellt, dann listet es alle möglichen ersten Wörter des Satzes auf, wie dies in Fig. 8 A angegeben ist. Nach Sprechen des ersten Wortes wird auf dem Schirm das festgestellte Wort angezeigt und es werden alle möglichen zweiten Wörter gemäß Fig. 8B aufgelistet. Dies setzt sich fort bis zum Ende des Satzes, wenn die Daten für eine Übertragung über dem Extremschmal-

band Nachrichtenkanal zusammengesetzt werden. Der Sprecher kann mit der Zeit sehen, welche nächsten Wörter erwartet werden. Der Rechner 20 überwacht die Genauigkeit der Wortübereinstimmungen. Fällt irgendein Wort unter einen adaptiven Schwellenwert, dann wiederholt die Synthetisiereinheit 22 den Satz und fragt nach fizierung vor der Durchführung. Werden alle Wörter ganz klar erkannt, dann gibt die Synthetisiereinheit22 den Satz nach Vervollständigung als Echo wieder, während der Rechner die Nachricht aussendet.

Nach Verarbeitung jedes gesprochenen Wortes wird dieses in den Speicher im Rechner 20 gebracht, wo die gesamte Nachricht in ein Digitalsignal für eine minimale oder fast minimale Anzahl von Bits codiert wird. Die Wörter können in codierter Form gespeichert werden, so daß sich der erforderliche Speicherplatz reduziert. Da das System eine vorbestimmte Anzahl von Wörtern enthält, die es erkennen kann, d.h., eine vorbestimmte Anzahl von Wortmodellen oder Mustern, so kann die Codierung in einer speziellen Nummer für jedes der Wörter bestehen. So kann im Beispiel der Fig. 8 den Wörtern "shift focus " die Nr. 12 und dem Wort "south" die Nr. 18 zugeordnet werden, während die Ziff. 2 durch die Nummer 21 dargestellt wird usw. Da diese Wörter durch die gleichen Nummern in dem entfernten Terminal 12 dargestellt werden, wandetl der Personalrechner 20 diese Nummern in ein Digitalsignal um und überträgt das Signal zu dem entfernten Terminal 12, wo das Signal in Nummern und dann in Wörter zurückgewandelt wird.

Ein zweites Codierungsverfahren, das bei dem vorliegenden Ausführungsbeispiel angewandt wird, besteht darin, jeden Buchstaben jedes Wortes in der ASC II-Codierung zu codieren. Dieses Codierungsverfahren hat einige Vorteile, obwohl es einige wenige Bits mehr pro Wort benötigt. Einer dieser Vorteile besteht darin, daß das ausgesandte Signal direkt zu den meisten heutigen elektrisch arbeitenden Druckvorrichtungen übertragen werden kann. In der ASC

II Codierung wird jeder Buchstabe durch 8 Bits dargestellt. Wenn somit die Musternachricht der Fig. 8 (shift focus south 22 miles" ist, dann ist die für die Übertragung dieser Nachricht in der ASC II Codierung erforderliche Bitzahl gleich 260. Werden 20 Bits zur Beschreibung von Eigenschaften der Stimme des Sprechers verwendet und erfordern Synchronisationsfehlererkennung und Steuersignale weitere 30 Bits, dann ist die vollständige Nachricht etwa 310 Bits lang. Es ist somit möglich eine Nachricht mit einer Länge von etwa 4 Sekunden und mit 310 Bits, d.h., mit etwa 77 Bits pro Sekunde zu übertragen.

Wird wie zuvor beschrieben ein Codierungssystem verwendet, bei dem jedem Wort eine spezielle Nummer zugeteilt ist, dann ist die Situation folgende: nimmt man an, daß die gesprochene Nachricht eine von 100 möglichen Nachrichtentypen mit jeweils gleicher Wahrscheinlichkeit ist, dann sind 7 Bits erforderlich um, um den gramatikalischen Aufbau der Nachricht zu beschreiben. Werden 20 auswählbare Wörter in dem System gespeichert die ausgewählt werden können, um verschiedene Positionen in der Nachricht einzunehmen, dann definieren 8 Bits welches Wort in jeder gewünschten Position in der Nachricht verwendet wurde. Für die Musternachricht, wie Sie zuvor angegeben wurde, nämlich für "shift focus south 22 miles" definieren 7 Bits die Nachricht Syntax, 40 Bits definieren die 5 auswählbaren Wörtern an Positionen innerhalb der Nachricht, wo eines von mehreren Wörtern ausgewählt werden kann, und etwa 20 Bits können die Eigenschaften der Stimme der Sprecher angeben, so daß sich eine Gesamtzahl von 67 Bits ergibt. Werden wiederum etwa 30 Bits für die Synchronisationsfehlerkorrektur und Steuersignale angesetzt, dann umfaßt die gesamte Nachricht etwa 97 Bits oder etwa 25 Bits pro Sekunde.

Die Synthetisiereinheit 22 des vorliegenden Ausführungsbeispiels ist im Handel erhältlich und wird von der Firma

Mikromint Inc. als Mikrovoxsynthesizer vertrieben. Der Fachmann erkennt selbstverständlich, daß die LPC-Analysiereinheit 15 einen Synthetisierer aufweist, (vgl. Fig. 2) und an Stelle der Synthetisiereinheit 22 verwendet, wenn die Sprechererkennung in dem System eingeschlossen ist und wenn es erwünscht ist, daß die syntetisierte Stimme der Stimme des ursprünglichen Sprechers gleicht. Die Synthetisiereinheit 22 wurde jedoch hier beschrieben und zwar der Einfachheit und des besseren Verständnis halber. Von der Beschreibung der Synthetisiereinheit 22 ergibt sich für den Fachmann ein vollkommenes Verständnis der Arbeitsweise des in der LPC-Analysiereinheit 15 vorhandenen Synthetisierers. Eine vollständigere Beschreibung des Synthetisierers der in der LPC-Analysiereinheit 15 enthalten ist, ergibt sich aus der zuvor genannten Patentanmeldung und aus der US-Patentanmeldung mit der Bezeichnung "Speech Synthesizer With Smooth Linear Interpolation", mit dem Aktenzeichen 267 203, eingereicht am 26 Mai 1981.

Die Synthetisiereinheit 22 ist ein freistehender inteligenter Mikroprozessor, der ASCII Text in gesprochenes Englisch umwandelt. Sie besteht aus einem M 65 02 Mikroprozessor 55, einer 9600 BPS UART-Teinheit 57 als serielle Schnittstelle einem RAM-Speicher 59 mit einer Speicherkapazität von 2K Bits einem löschbaren, programmierbaren Nur-Lesespeicher EPROM 61 mit 8 K-Bits, einem SC01 Votrax-Stimmsyntetisierer 63, einem taktenden und programmierbaren Teiler 65 und verschiedenen Puffern, Steuerungen und Verstärkern. Die Synthetisiereinheit 22 verwendet einen Algorythmus, der grammatikalisch Serieneingangsdaten in Wörter umsetzt, dann die englischen Ausspracheregeln verwendet und einen Lautstrom aus dem ausgesprochenen zu erzeugen. Dieser Lautstrom steuert dann den Sprachsynthetisierer 63. Der Sprachsynthetisierer 63 besitzt einen ROM-Speicher der Laute als eine Folge von 1 bis 4 Tönen in stätigem Zustand von spezifischer Dauer und mit spezifischem Spektrum erzeugt. Die Funktion der Synthetisiereinheit 22 beruht auf den Buchstaben zu Laut-Umsetzungsregeln, die

in dem Mikroprozessor 55 angewandt werden, sowie auf der Laut-Sprachensyntese in dem Sprachsynthetisierer 63. Der Mikroprozessor 55 liest bis zu 1500 Zeichen in seinen internen Seitenpuffer von der seriellen Schnittstelle 57. Er identifiziert Phrasengruppen durch ihre Punktuation und Wörter durch ihre Zwischenraumbegrenzer. Er verwendet die Phrasengruppengrenzen um eine geeignete deklarative oder fragende Tonhöhen- und Dauerbeugung auf die Phrase anzuwenden. Pro Wort wird jedes Zeichen von links nach rechts über das Wort abgetastet. Wird ein Zeichen gefunden, bei dem die linken und rechten Kontexterfordernisse, (benachbarte Zeichen) erfüllt sind, dann wird die erste anwendbare Regel für das Zeichen verwendet, um es in einen Laut umzusetzen.

Der Sprachsynthetisierer 63 ist ein CMOS-Typ, der aus einem digitalen Codeumsetzer und einem elektronischen Modell der Vokalspur besteht. Intern ist eine Lautsteuerung vorgesehen, die eine 6-Bit-Laut- und 2-Bit-Tonhöhencodierung in eine Matrix von spektralen Parametern umsetzt, die das Vokalspurmodell zur Synthetisierung der Sprache einstellt. Die Ausgangstonhöhe der Laute wird durch die Frequenz des gegetakteten Teiler 65 abgegebenen Taktsignal gesteuert. Feine Schwankungen der Tonhöhe können induziert werden, um eine Beugung hinzuzufügen, was verhindert, daß die synthetisierte Stimme monoton und maschinell klingt. Während der vorliegende Algorrhythmus einen englischen Text in Sprache umwandelt, ist es für den Fachmann verständlich, daß die Sprachalgorhythmus genauso in anderen Sprachen geschrieben sein können. 64 Laute definieren die englische Sprache und jeder Laut wird durch eine 6-Bit-Codierung gekennzeichnet, die von dem Mikroprozessor 55 an den Sprachsynthetisierer 63 angelegt wird. Die Lautsteuerung setzt dann die Bits in die zuvor erwähnten Spektralparameter um.

Damit die synthetisierte Sprache möglichst gut dem identifizierten ursprünglichen Sprecher gleicht, können verschiedene Codierungen senderseitig zu dem empfangenden Gerät

übertragen werden, wobei Daten über die spezielle Aussprache des Sprechers bezüglich dieser Worte beinhalten. Dies kann sehr einfach dadurch erreicht werden, daß eine Sprecheridentifikationscodierung ausgesandt wird, die der Empfänger zum Aufsuchen der Vokalspurlänge und des mittleren Tonhöhenbereichs verwendet. Alternativ dazu kann der Sender auch Polynomkoeffizienten aussenden, die die Tonhöhenkontur über der Länge des Satzes beschreibt, sowie einen Vokalspurlängenmodifizierer. Diese Polynomkoeffizienten ermöglichen, daß der richtige Tonhöhenbereich, Tonhöhenabfall und die Betonung mit sehr wenigen Bits übertragen werden. Der Vokalspurlängenmodifizierer ermöglicht es dem Syntetisierer eine Polynominterpolation der LPC-Reflektionskoeffizient durchzuführen, wodurch die Vokalspur länger oder kürzer gemacht werden kann als bei dem gespeicherten Muster, das bei den Buchstaben- Zu- Ton-Regeln verwendet wird.

Es wurde somit ein Extremschmalband-Übertragungssystem offenbart, bei dem jedes Terminal menschliche Stimme in Digitalsignale mit einer Geschwindigkeit von weniger als 300 Bits pro Sekunde umsetzt. Das Terminal besitzt ferner die Fähigkeit Digitalsignale zu empfangen, die repräsentativ für eine menschliche Stimme sind, und die menschliche Stimme mit den gleichen Eigenschaften wie die des ursprünglichen Sprechers zu synthetisieren. Außerdem besitzt jedes Terminal die Fähigkeit Wörter und den speziellen Sprecher mit sehr hoher Genauigkeit zu erkennen.

– Leerseite –

Num[illegible] 34 16 238
Int. C[illegible] G 10 L 1/00
Anmeldetag: 2. Mai 1984
Offenlegungstag: 20. Dezember 1984

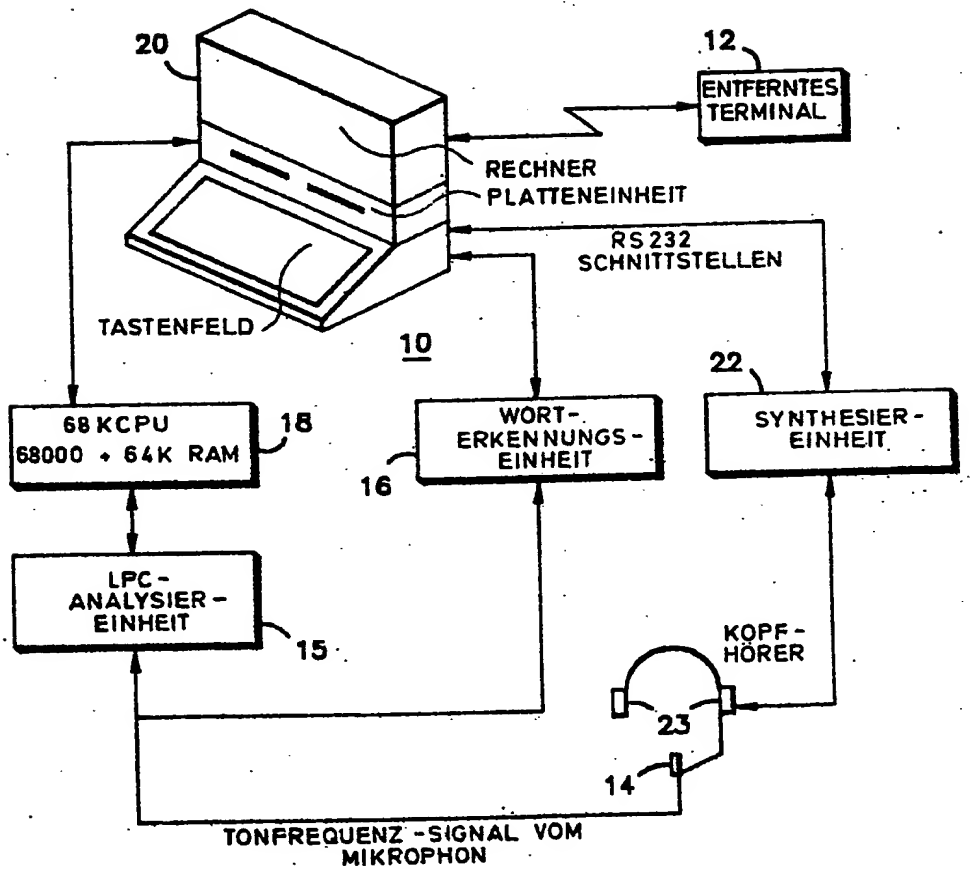


FIG. 1

NACHGEREICHT

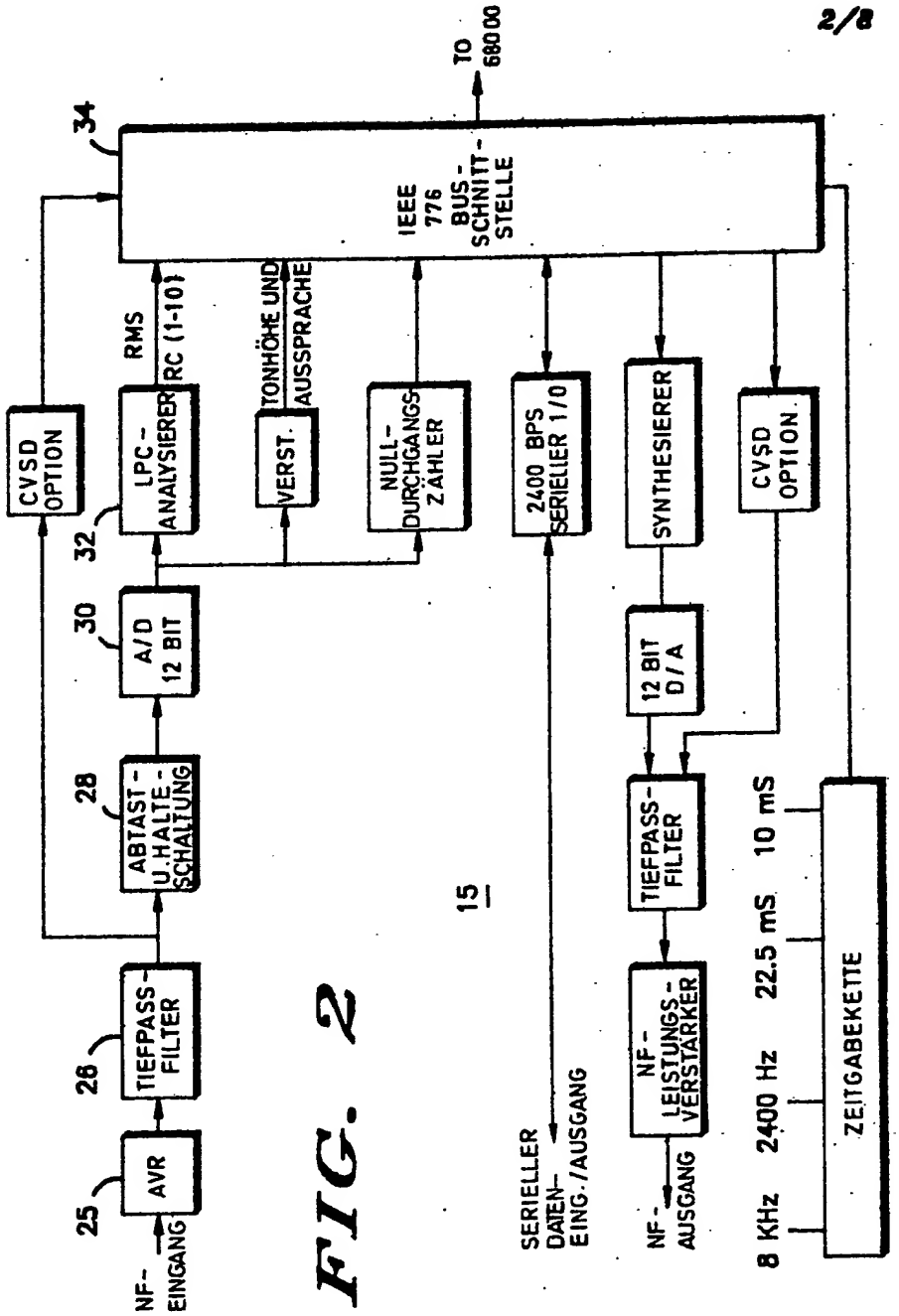
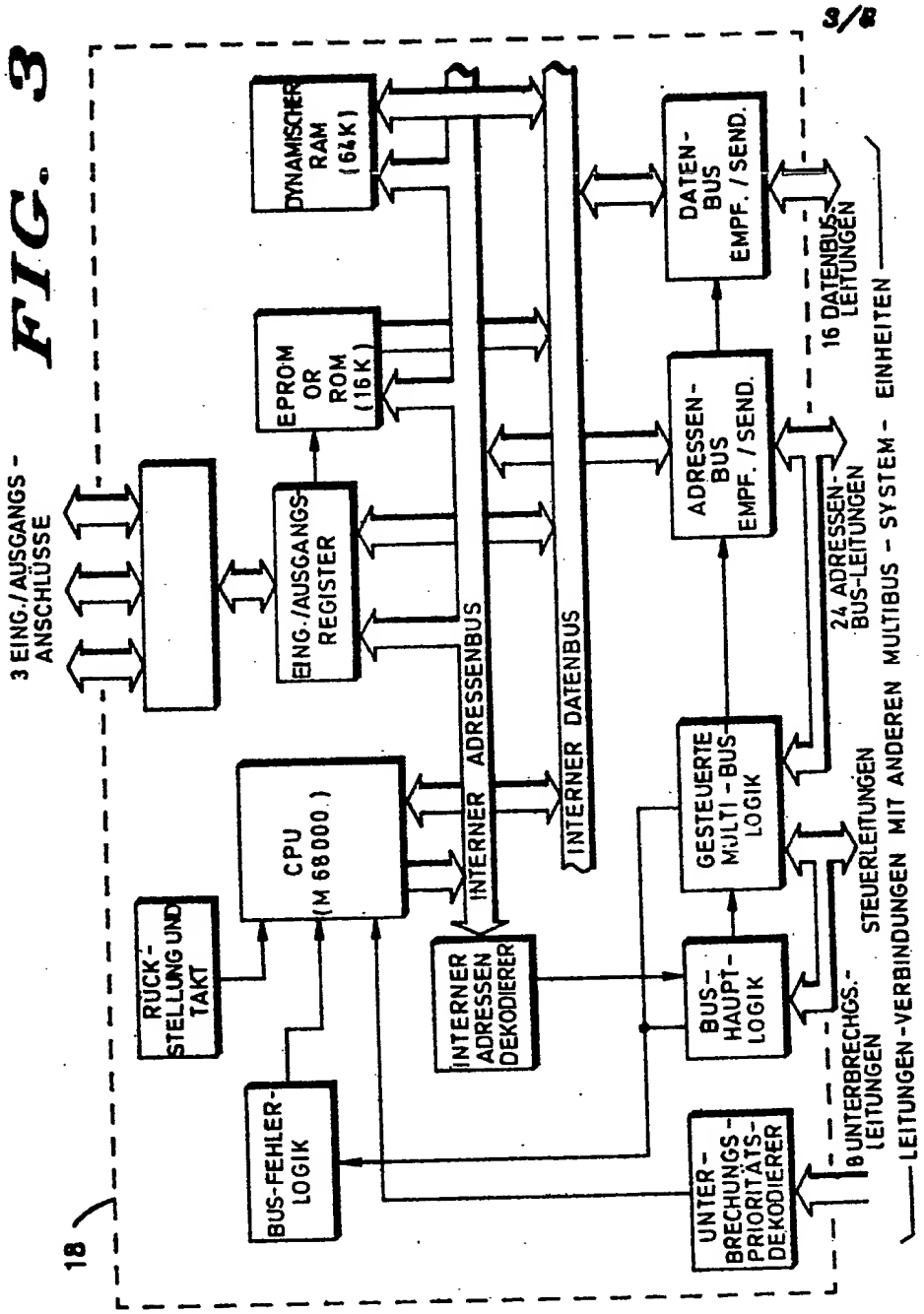


FIG. 2

FIG. 3



3416238

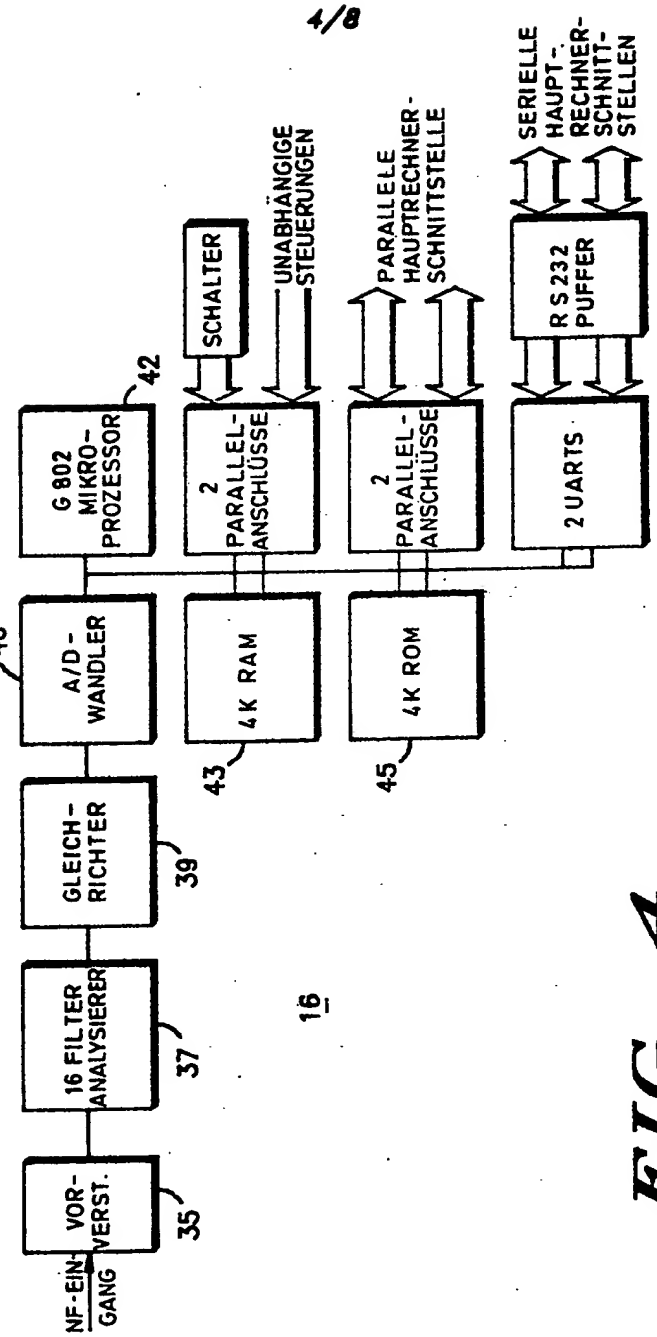
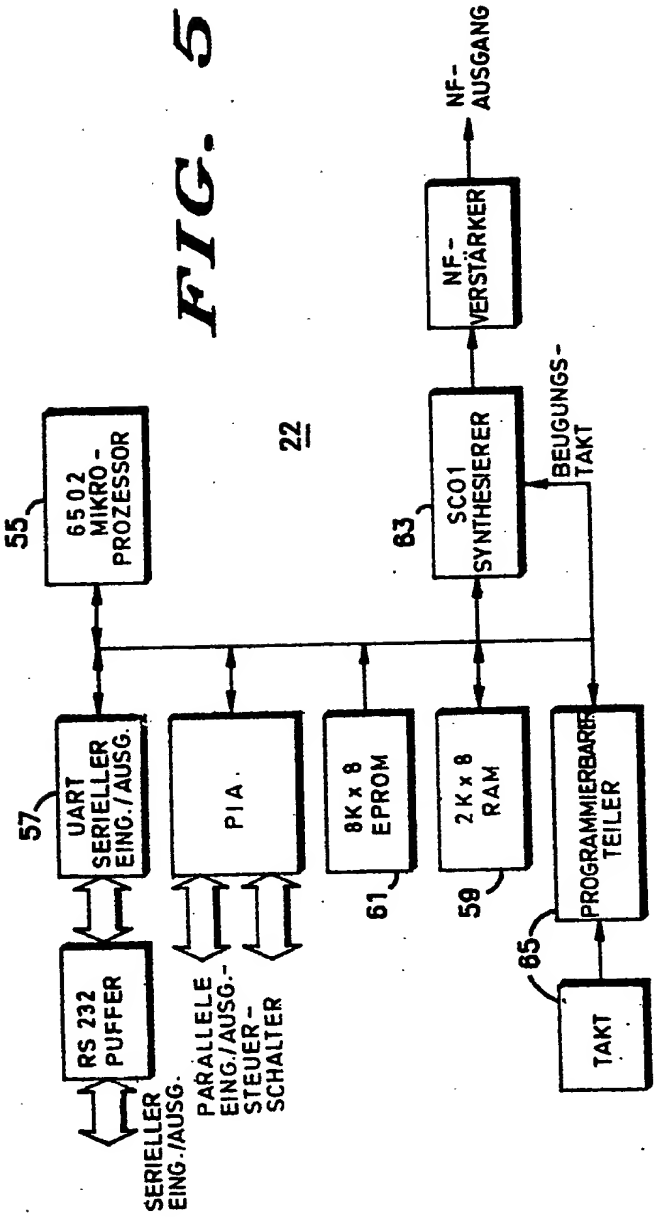


FIG. 4

3416238

FIG. 5



3416238

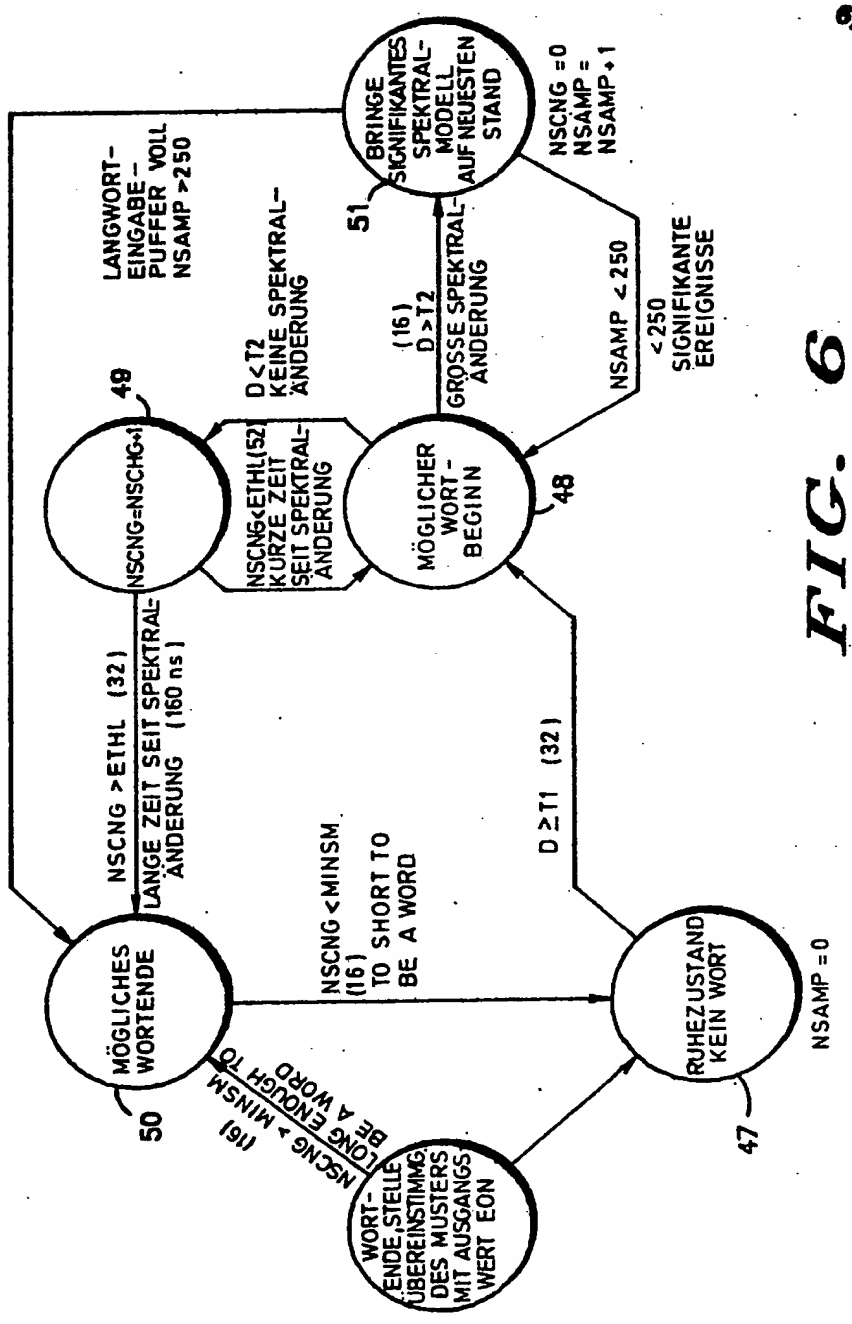
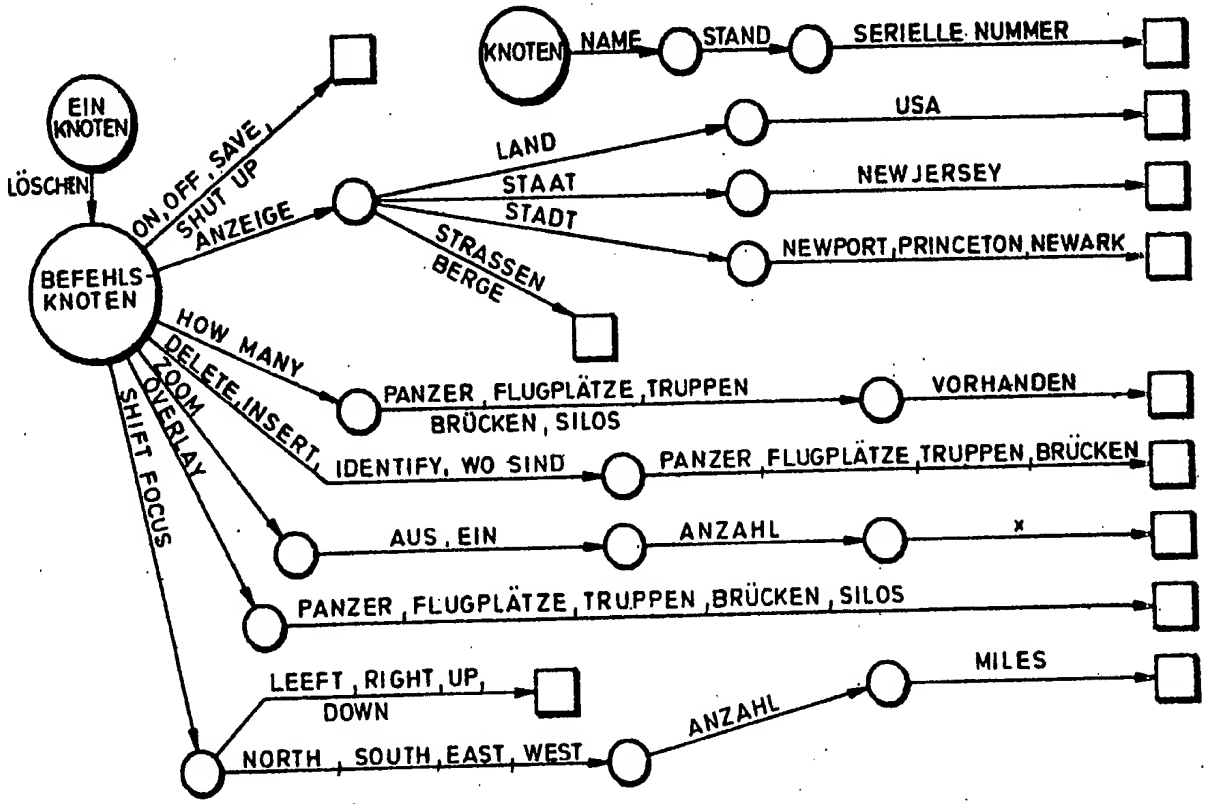


FIG. 6

FIG. 7

BEMERKUNGEN: 1. LÖSCHEN KANN BEI JEDEM KNOTEN VERWENDET WERDEN, AUSSER BEI LETZTEM KNOTEN-, UM EINEN KNOTEN IM ENTSCHEIDUNGSBAUM ZU ERSETZEN
2. "UH" UND "THE" WERDEN IMMER IGNORIERT



30

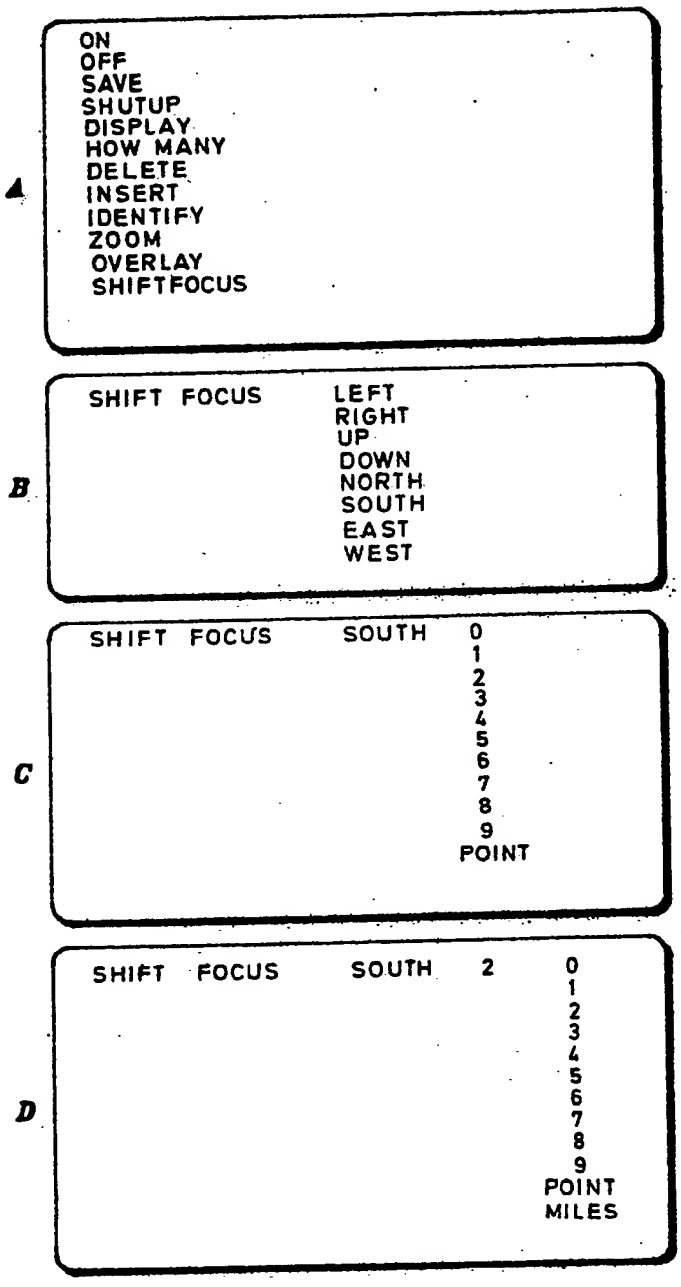


FIG. 8